Exploiting Auditory Fovea in Humanoid-Human Interaction

Kazuhiro Nakadai[†], **Hiroshi G. Okuno**[†*], **and Hiroaki Kitano**[†‡]

†Kitano Symbiotic Systems Project, ERATO, Japan Science and Technology Corp.
Mansion 31 Suite 6A, 6-31-15 Jingumae, Shibuya-ku, Tokyo 150-0001, Japan
Tel: +81-3-5468-1661, Fax: +81-3-5468-1664
* Department of Intelligence Sciece and Technology, Graduate School of Informatics, Kyoto University
‡Sony Computer Science Laboratories, Inc.
nakadai@symbio.jst.go.jp, okuno@nue.org, kitano@csl.sony.co.jp

Abstract

A robot's auditory perception of the real world should be able to cope with motor and other noises caused by the robot's own movements in addition to environment noises and reverberation. This paper presents the *active direction-pass filter* (ADPF) that separates sounds originating from a specified direction detected by a pair of microphones. Thus the ADPF is based on directional processing – a process used in visual processing. The ADPF is implemented by hierarchical integration of visual and auditory processing with hypothetical reasoning of interaural phase difference (IPD) and interaural intensity difference (IID) for each sub-band. The ADPF gives differences in resolution in sound localization and separation depending on where the sound comes from: the resolving power is much higher for sounds coming directly from the front of the humanoid than for sounds coming from the periphery. This directional resolving property is similar to that of the eye whereby the visual fovea at the center of the retina is capable of much higher resolution than is the periphery of the retina. To exploit the corresponding "auditory fovea", the ADPF controls the direction of the head. The human tracking and sound source separation based on the ADPF is implemented on the upper-torso of the humanoid and runs in real-time using distributed processing by 5 PCs networked via a gigabit ethernet. The signal-to-noise ratio (SNR) and noise reduction ratio of each sound separated by the ADPF from a mixture of two or three speeches of the same volume were increased by about 2.2 dB and 9 dB, respectively.

Introduction

Robots are often used as a real-world test-beds in the research fields of AI and cognition. Robots can serve as remote communication agents with a real body. Such a robot can provide remote conferencing of high quality than conventional TV conferencing by enabling listeners to hear several things simultaneously or to keep attention on a specific speaker. The robot achieves this by providing sound with a high signal-to-noise ratio (SNR) obtained by suppressing unwanted sounds from other speakers and by suppression of the noises from the robot's own motors.

Therefore, a remote robot should be able to cope with a general sound, i.e. a mixture of sounds. To localize and separate sound sources accurately, the robot needs to be equipped with two or more microphones and the difference in resolution, according to the direction of the sound, needs to be taken into account.

When a robot has a pair of microphones in ear positions, the sensitivity of sound source localization in the azimuth is the highest to the front of the head, and degrades towards the periphery. In humans, this phenomenon has been known for over a century (Blauert 1999). Neurons of horseshoe bats are narrowly tuned in a specialized region inside the cochlea to detect frequency shifts in the echo signal that are caused by Doppler-effects. In neuroethology, then this is termed an "auditory fovea" corresponding to the visual fovea in the primate retina (Schuller & Pollak). The concepts of the phenomena in sound source localization by the robot and the cochlea in horseshoe bats are similar to visual fovea in the primate retina in a point of selective attention. So, we can say that the phenomenon in sound source localization by the robot is a kind of auditory fovea. In this paper, we use an auditory fovea in a sense of higher sensitivity in front direction of the robot's head.

In the retina, the resolution of images is high in the fovea, which is located at the center of the retina, and much poorer towards the periphery which serves to capture information from a much larger area. Because the visual fovea gives a good compromise between the resolution and field-of-view without the cost of processing a large amount of data, it is useful for robots (Klarquist & Bovik 1998; Rougeaux & Kuniyoshi 1997). The visual fovea must face the target object to obtain good resolution, so it is a kind of *active vision* (Aloimonos, Weiss, & Bandyopadhyay. 1987). Akin to the visual fovea, the auditory fovea also needs to be directed at the target object, such as a speaker. Therefore, it too relies on active motion. Such integration of sound and active motion, termed *active audition* (Nakadai *et al.* 2000b), can be used to attain improved auditory perception. The active motion is essential in audition and vision not only for friendly humanoid-human interaction, but also for better perception.

Active audition has been integrated with multiple face recognition and visual localization. A real-time multiple human tracking system has been developed (Nakadai *et al.*

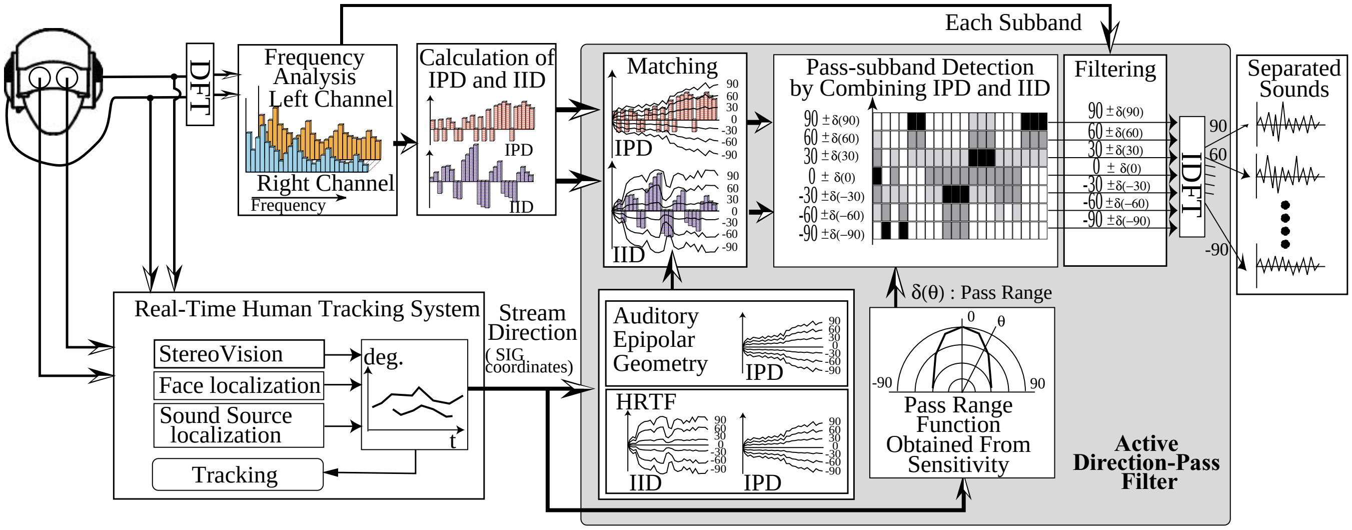

Figure 1: Active Direction-Pass Filter

2001). The system can even simultaneously track more than one voice, owing to robust auditory processing of harmonic structures of sound and time-streamed audio-visual integration.

However, the system has drawbacks:

1. Localization is attained but without sound source separation or enhancement.
2. Multiple sound sources are localized by a simple grouping strategy using harmonic structure of sounds.
3. The accuracy of sound source localization is not high enough for sound source separation.
4. No visual information is used when a face is not in sight.
5. The system network is difficult to scale-up.

To cope with the first and second issues, we propose a new sound source separation method called the *active direction-pass filter (ADPF)*. It is a kind of direction pass filter (DPF) that extracts a sound from a specific direction with hypothetical reasoning about the interaural phase difference (IPD) and interaural intensity difference (IID) of each sub-band (Okuno *et al.* 2001). It enables more accurate sound source extraction by auditory fovea based separation, active motion, and accurate localization by audio-visual integration. Other common techniques of sound source separation are a beam forming with a microphone array (Asano, Asoh, & Matsui 1999), ICA as blind source separation (Okuno, Ikeda, & Nakatani 1999), and computational auditory scene analysis (CASA) techniques for understanding general sounds. However, these techniques assume that assumes that the microphone setting is fixed. Some techniques assume that the number of microphones is more than or equal to the number of sound sources – an assumption that may not hold in a real-world environment. Asano *et al.* also reported a robot which separates and localizes sound sources by using a 8 ch circle microphone array in an ordinary office room (Asano *et al.* Sep 2001). However, their system requires a lot of measurement for separation and localization in advance, and has difficulty in sound source separation during motion, while human can hear during motion. It is not enough for robot audition to be deployed in the real world yet. A sound source separation method based on audio-visual integration has been reported (Nakagawa, Okuno, & Kitano 1999). However, it uses only simple visual clues for integration and works only in off-line and simulated environments because the data was created by convolution of a *head related transfer function (HRTF)* measured in an anechoic room.

For the third and fourth issues, stereo vision is introduced for accurate and robust localization even when a person looks away. In addition, a Kalman filter is implemented for reducing measurement and process noises in localization.

For the last issue, because a PC is added to the system for stereo vision, a more scalable network is required. In our network, gigabit and fast ether were combined to improve scalability. In addition, accurate synchronization between PCs is given by using a network time protocol (NTP) and an original synchronization protocol. To summarize, in this paper, we describe the following achievements:

1. sound source separation by using an auditory fovea
2. accurate stream formation by a Kalman filter
3. accurate localization by stereo vision
4. scalable network communication by the NTP and the synchronization protocol.

The rest of this paper is organized as follows: Section 2 describes the active direction-pass filter. Section 3 explains the real-time human tracking system refined by the Kalman filter and stereo vision. Section 4 evaluates the performance of the ADPF. The last sections provide a discussion and a conclusion.

Active Direction Pass Filter

The architecture of the ADPF is shown in Figure 1. The ADPF separates sound sources from four inputs – a spec-

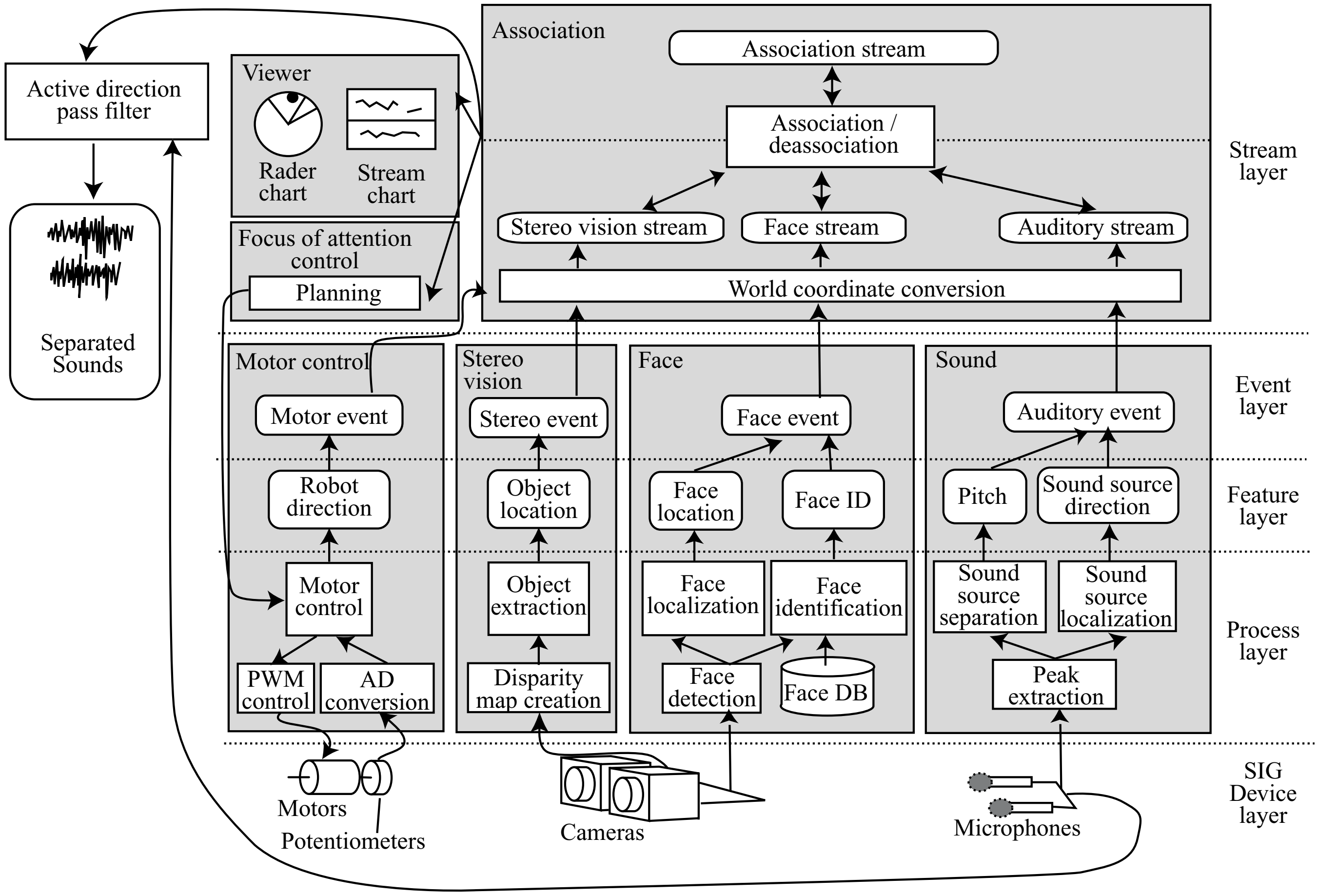

Figure 2: Hierarchical Architecture of Real-Time Tracking System

trum of input sound, interaural intensity difference (IID), interaural phase difference (IPD) and sound source directions. The spectrum is obtained from sound captured by the robot's microphone. The IPD and IID are calculated from spectra of the left and right channels. Sound source directions are obtained from streams generated in the real-time human tracking system as described in a later section.

The APDF uses two key techniques, auditory epipolar geometry and auditory fovea. The auditory epipolar geometry is a localization method by IPD without using HRTFs. The auditory epipolar geometry is described in the next section in detail. In this paper, the ADPF is implemented to use both HRTFs and auditory epipolar geometry for evaluation. The auditory fovea is used to control the pass range of the ADPF: the pass range is narrow in the front direction and wider in the periphery. The detailed algorithm of the ADPF is described below:

1. IPD $\Delta\varphi'$ and IID $\Delta\rho'$ in each sub-band are obtained by the difference between the left and right channels.
2. Let θ_s be the azimuth of a stream with current attention in the robot coordinate system in the real-time human tracking system. θ_s is sent to the ADPF through the gigabit ether network with consideration of the latency of the processing.
3. The pass range $\delta(\theta_s)$ of the ADPF is selected according to θ_s. The pass range function δ has a minimum value in the *SIG* front direction, because then it has maximum sensitivity. δ has a larger value at the peripheral because of the lower sensitivity. Let us $\theta_l = \theta_s - \delta(\theta_s)$ and $\theta_h = \theta_s + \delta(\theta_s)$.
4. From a stream direction, the IPD $\Delta\varphi_E(\theta)$ and IID $\Delta\rho_E(\theta)$ are estimated for each sub-band by the auditory epipolar geometry. Likewise, the IPD $\Delta\varphi_H(\theta)$ and IID $\Delta\rho_H(\theta)$ are obtained from HRTFs.
5. The sub-bands are collected if the IPD and IID satisfy the specified condition. Three conditions are applied:

   A: $f < f_{th}$ **:** $\Delta\varphi_E(\theta_l) \le \Delta\varphi' \le \Delta\varphi_E(\theta_h)$,
   B: $f < f_{th}$ **:** $\Delta\varphi_H(\theta_l) \le \Delta\varphi' \le \Delta\varphi_H(\theta_h)$, and
   $f \ge f_{th}$ **:** $\Delta\rho_H(\theta_l) \le \Delta\rho' \le \Delta\varphi_H(\theta_h)$,
   C: $f < f_{th}$ **:** $\Delta\varphi_E(\theta_l) \le \Delta\varphi' \le \Delta\varphi_E(\theta_h)$, and
   $f \ge f_{th}$ **:** $\Delta\rho_H(\theta_l) \le \Delta\rho' \le \Delta\varphi_H(\theta_h)$.

   f_{th} is the upper boundary of frequency which is efficient for localization by IPD. It depends on the baseline of the ears. In *SIG*'s case, the f_{th} is 1500 Hz.
6. A wave consisting of collected sub-bands is constructed.

Note, that to obtain a more accurate direction, the direction of an association stream is specified by visual information not by auditory information.

Auditory Epipolar Geometry

HRTFs obtained by measurement of a lot of impulse responses are often used for sound source localization in binaural research. Because HRTF is usually measured in an anechoic room, sound source localization in an ordinary echoic room needs HRTF including room acoustic, that is, the measurement has to be repeated if the system is installed at different room. However, deployment to the real world means that the acoustic features of the environment are not known in advance. It is infeasible for any practical system to require such extensive measurement of the operating space. Thus, audition system without or at least less dependent on HRTF is essential for practical systems.

Auditory Epipolar Geometry can extract directional information of sound sources without using HRTF (Nakadai *et al.* 2000a). In stereo vision research, epipolar geometry is one of the most commonly used localization methods (Faugeras 1993). Auditory epipolar geometry is an extension of epipolar geometry in vision (hereafter, *visual epipolar geometry*) to audition. Since auditory epipolar geometry extracts directional information geometrically, it can dispense with HRTF. However, the reported auditory epipolar geometry does not take the effect of the cover into account and so it must be refined. The refined auditory epipolar geometry works as follows:

First, for each sub-band, it calculates the IPD from a pair of spectra obtained by *fast fourier transform* (FFT). Then, the sound source direction is estimated by

$$\theta = D^{-1}\left(v/(2\pi f)\Delta\varphi\right), \tag{1}$$

where D represents the difference between the distances of the left and right ears from a sound source, v is the velocity of sound, f is the frequency of sound and $\Delta\varphi$ is *IPD*. In this paper, the velocity of sound is fixed at 340m/sec, irrespective of temperature and humidity.

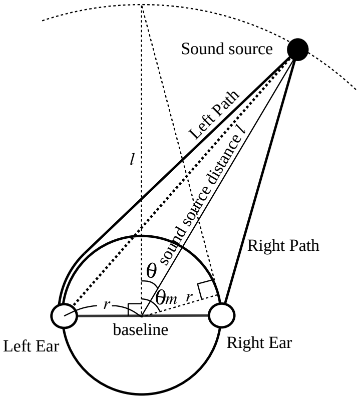

Figure 3: Auditory Epipolar Geometry

On defining D, the influence of the cover of *SIG* should be considered. The cover prevents sound from reaching the ears directly. The sound in Figure 3, for example, has to travel along the cover because the path between the left ear and the sound source is actually not direct. The problem is solved by adjusting the formula for auditory epipolar geometry by taking the shape of *SIG* into account. The formulae are specified as follows:

$$D(\theta,l) = \begin{cases} r\left(\pi - \theta - \theta_m\right) + \delta(\theta,l) & \left(0 \le \theta < \frac{\pi}{2} - \theta_m\right) \\ r\left(\pi - 2\theta\right) & \left(|\theta - \frac{\pi}{2}| \le \theta_m\right) \\ r\left(\theta - \theta_m\right) + \delta(\pi - \theta,l) & \left(\frac{\pi}{2} + \theta_m < \theta \le \pi\right) \end{cases} \tag{2}$$

$$\delta(\theta,l) = \sqrt{l^2 - r^2} - \sqrt{l^2 + r^2 - 2rl\cos\theta}, \tag{3}$$

$$\theta_m = \arcsin\frac{r}{l}. \tag{4}$$

Thus, D is defined as a function of θ and l. Figure 4 shows the relationship of D, θ and l obtained by simulation. The larger the θ, the bigger the influence of l. However, when l is more than 50 cm, the influence of l can be ignored. In such a case, we can take l to be infinite and define D as a function of only θ as follows:

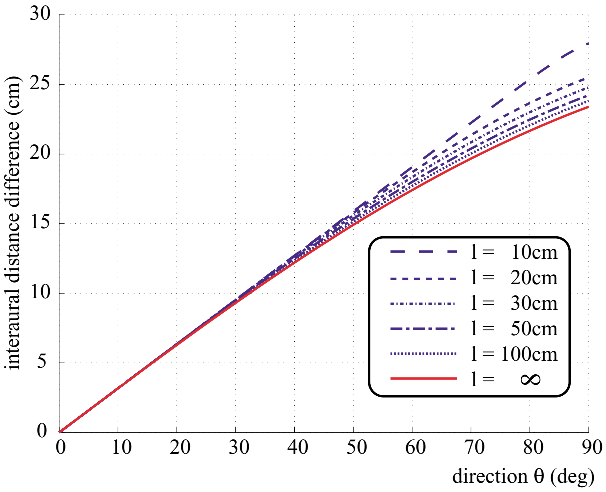

Figure 4: IPD and distance from sound source

$$\begin{aligned} D(\theta) &= \lim_{l\to\infty} D(\theta,l) \\ &= r\left(\theta + \sin\theta\right). \end{aligned} \tag{5}$$

Since the baselines for vision and audition are in parallel in *SIG*, whenever a sound source is localized by visual epipolar geometry, it can be easily converted into the angle θ. This means that a symbolic representation of direction is used as a clue for the integration of the visual and auditory information. We have reported the feasibility of such an integration based on epipolar geometry (Nakadai *et al.* 2000a).

Real-Time Human Tracking System

The following sections describe other components associated with the ADPF – the humanoid *SIG* and the real-time human tracking system. They provide input to the ADPF.

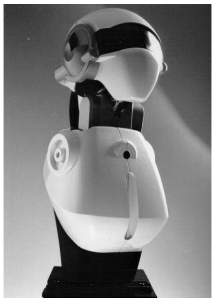

Figure 5: Humanoid *SIG*

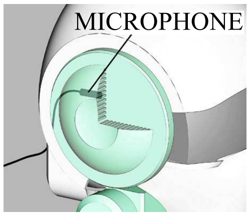

Figure 6: *SIG* microphone

Compared with the reported system (Nakadai *et al.* 2001), the real-time tracking system has been improved by tracking based on a Kalman filter which gives a stronger directional property.

Platform: SIG

The upper torso humanoid *SIG*, shown in Figure 5, is a testbed for humanoid-human interaction . *SIG* has a fiber reinforced plastic (FRP) cover designed to acoustically separate the *SIG* inner and external worlds. A pair of CCD cameras (Sony EVI-G20) is used for stereo vision. Two pairs of microphones are used for auditory processing. One pair is located in the left and right ear positions for sound source localization (Figure 6). The other pair is installed inside the cover, mainly for canceling noise from the robot's motors. *SIG* has 4 DC motors (4 DOFs), the position and velocity of which are controlled through potentiometers.

The Real-Time Human Tracking System

The architecture of the real-time human tracking system is shown in Figure 2. It consists of seven modules, Sound, Face, Stereo Vision, Association, Focus-of-Attention, Motor Control and Viewer.

Sound localizes sound sources. Face detects multiple faces by combining skin-color detection, correlation based matching, and multiple scale image generation (Hidai *et al.* 2000). It identifies each face by Linear Discriminant Analysis (LDA), which creates an optimal subspace to distinguish classes and continuously updates the subspace on demand with a little of computation (Hiraoka *et al.* 2000). In addition, the faces are localized in 3-D world coordinates by assuming an average face size. Stereo Vision is a new module to localize precisely lengthwise objects such as people precisely by using fast disparity map generation (Kagami *et al.* 1999). It improves the robustness of the system by being able to track a person who looks away and does not talk. Association forms *streams* and associates them into a higher level representation, that is, an *association* stream according to proximity. The directions of the streams are sent to the ADPF with captured sounds. Focus-of-Attention plans *SIG*'s movement based on the status of streams. Motor Control is activated by the Focus-of-Attention module and generates pulse width modulation (PWM) signals to the DC motors. Viewer shows the status of auditory, visual and association streams in the radar and scrolling windows. The whole system works in real-time with a small latency of 500 ms by distributed processing with 5 PCs, networked through gigabit and fast ethernet.

Stream Formation and Association: Streams are formed in Association by connecting events from Sound, Face and Stereo Vision to a time course.

First, since location information about sound, face and stereo vision events is observed in a *SIG* coordinate system, the coordinates are converted into world coordinates by comparing a motor event observed at the same time.

The converted events are connected to a stream through a Kalman filter based algorithm. The Kalman filter efficiently reduces the influence of process and measurement noises in localization, especially in auditory processing with bigger ambiguities.

In Kalman filter based stream formation, position $\boldsymbol{p}$ with a dimension N is approximated by a recursive equation defined by

$$\begin{aligned} \boldsymbol{p}_{k+1} &= \boldsymbol{p}_k + \boldsymbol{v}_k \Delta T \\ &= \boldsymbol{p}_k + (\boldsymbol{p}_k - \boldsymbol{p}_{k-l})/l, \end{aligned} \tag{6}$$

where l is a parameter for average velocity.

When $\boldsymbol{x}_k$ is a state vector represented as $(\boldsymbol{p}_k, \boldsymbol{p}_{k-1}, \cdots, \boldsymbol{p}_{k-l})$ and $\boldsymbol{y}_k$ is a measurement represented as a position vector, and functions to estimate the state and measurement of the process are defined by

$$\begin{aligned} \boldsymbol{x}_{k+1} &= F\boldsymbol{x}_k + G\boldsymbol{w}_k, \\ \boldsymbol{y}_k &= H\boldsymbol{x}_k + \boldsymbol{v}_k, \end{aligned} \tag{7}$$

where $\boldsymbol{w}_k$ and $\boldsymbol{v}_k$ represent the process and measurement noise, respectively. F, G and H are defined as follows:

$$F = \left(\begin{array}{cccc|c} \frac{l+1}{l}I_N & \mathbf{0} & \cdots & \mathbf{0} & -\frac{1}{l}I_N \\ \hline I_N & & & \mathbf{0} & \\ & & \ddots & & \mathbf{0} \\ \mathbf{0} & & & I_N & \end{array}\right), \tag{8}$$

$$\begin{aligned} G &= (I_N \;\; \mathbf{0} \;\; \cdots \;\; \mathbf{0})^T, \\ H &= (I_N \;\; \mathbf{0} \;\; \cdots \;\; \mathbf{0}), \end{aligned} \tag{9}$$

where I_N is the identity matrix of $N \times N$ dimensions.

Then, the Kalman filter is defined as follows:

$$\begin{aligned} \hat{x}_{k|k} &= \hat{x}_{k|k-1} + K_k(y_k - H\hat{x}_{k|k-1}), \\ \hat{x}_{k+1|k} &= F x_{k|k}, \end{aligned} \tag{10}$$

$$K_k = \hat{P}_{k|k-1} H^T (I_N + H\hat{P}_{k|k-1}H^T)^{-1}, \tag{11}$$

$$\begin{aligned} \hat{P}_{k|k} &= \hat{P}_{k|k-1} - K_k H \hat{P}_{k|k-1}, \\ \hat{P}_{k+1|k} &= F\hat{P}_{k|k}F^T + \sigma_w^2/\sigma_v^2 GG^T, \\ \hat{x}_{0|-1} &= \bar{x}_0, \quad \hat{P}_{0|-1} = \sum x_0/\sigma_v^2, \end{aligned} \tag{12}$$

where $\hat{x}$ is an estimation of $\boldsymbol{x}$, K_k is the Kalman gain, $\hat{P}$ is an error covariance matrix. σ_w^2 and σ_v^2 are variance-covariance matrixes of $\boldsymbol{w}_k$ and $\boldsymbol{v}_k$.

An current position vector is estimated by

$$\hat{y}_k = H\hat{x}_{k|k}. \tag{13}$$

In sound stream formation, when a sound stream and an event have a harmonic relationship, and the difference in azimuth between $\hat{y}_k$ of the stream and a sound event is less than $\pm 10°$, they are connected.

In face and stereo vision stream formation, a face or a stereo stream event is connected to a face or a stereo vision stream when the difference in distance between $\hat{y}_k$ of the stream and the event is within 40 cm, and when they have the same event ID. An event ID is a face name or an object ID generated in face or stereo vision module.

When the system judges that multiple streams originate from the same person, they are associated into an association stream, a higher level stream representation. When one of the streams forming an association stream is terminated, the terminated stream is removed from the association stream, and the association stream is de-associated into one or more separated streams.

Control of Tracking: The tracking is controlled by Focus-of-Attention to keep the direction of a stream with attention and sends motor events to Motor. By selecting a stream with attention and tracking it, the ADPF can continue to make the best use of foveal processing The streams are separated, according to the surrounding situations, that is, the Focus-of-Attention control is programmable. In this paper, for the ADPF, the precedence of Focus-of-Attention control for an associated stream, including a sound stream, has the highest priority, a sound stream has the second priority and other visual streams have the third priority.

Performance Evaluation

The performance of the ADPF was evaluated by two kinds of experiments. In these experiments, *SIG* and loudspeakers of B&W Nautilus 805 were located in a room of 10 square meters. The *SIG* and loudspeakers were set at exactly the same height, 1 m apart shown in Figure 7. The number of loudspeakers depends on each experiment. The direction of the loudspeaker was represented as 0° when facing towards the *SIG*.

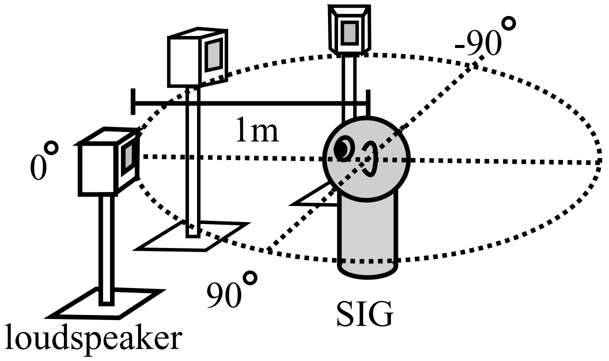

Figure 7: Conditions in the evaluation

For speech data, 20 sentences were read by men and women from the Mainichi Newspaper articles in ASJ Continuous Speech Corpus. Four kinds of metrics were used in the evaluation:

1. a difference in SNR in the frequency domain between the input and separated speech defined by
$$R_1 = 10\log_{10}\frac{\sum_{j=1}^{n}\sum_{i=1}^{m}(|sp(i,j)| - \beta|sp_o(i,j)|)^2}{\sum_{j=1}^{n}\sum_{i=1}^{m}(|sp(i,j)| - \beta|sp_s(i,j)|)^2}, \tag{14}$$
where $sp(i,j)$, $sp_o(i,j)$ and $sp_s(i,j)$ are the spectra of the original signal, the signal picked-up by the robot's microphones and the signal separated by the ADPF, respectively. m and n are the number of sub-bands and samples, respectively. β is the attenuation ratio of amplitude between the original and observed signals.
2. signal loss between the input and separated speech is defined by
$$R_2 = 10\log_{10}\frac{\sum_{n\in S}(s(n) - \beta s_o(n))^2}{\sum_{n\in S}(s(n) - \beta s_s(n))^2}, \tag{15}$$
where $s(n)$, $s_o(n)$ and $s_s(n)$ are the original signal, the signal picked by the robot's microphones and the signal separated by the ADPF, respectively. S is a set of samples with signals, that is, a set of i satisfying $s(i) - \beta s_o(i) \geq 0$.
3. effect of noise suppression is defined by
$$R_3 = 10\log_{10}\frac{\sum_{n\in N}(s(n) - \beta s_o(n))^2}{\sum_{n\in N}(s(n) - \beta s_s(n))^2}, \tag{16}$$
where $s(n)$, $s_o(n)$ and $s_s(n)$ are the same as Eq. (15). N is a set of samples with noises, that is, a set of i satisfying $s(i) - \beta s_o(i) < 0$.
4. evaluation by experts in audio signal processing.

Experiment 1: The errors of sound source localization of Sound, Face and Stereo Vision were measured with the sound source direction varied between 0° and 90°.

Experiment 2: The efficiency of the Kalman filter was measured. Two loudspeakers were used. One was fixed in the direction of 60°. The other was repeatedly moved from left to right within ±30°. Voices from the second loudspeaker were extracted by the ADPF. Two kinds of sound streams – with and without the Kalman filter – were used as inputs to the ADPF. The extracted sounds were compared by R_1.

Experiment 3: The efficiency of the ADPF by each filtering condition **A** to **C** described in the section "Active Direction Pass Filter" were measured by using the metrics R_1, R_2 and R_3. The separation of two or three simultaneous voices were assessed. The first loudspeaker was fixed at 0°. In the separation of two simultaneous voices, the second one was facing in a direction of 30°, 60° and 90° with respect to the *SIG*. In the separation of three simultaneous voices, the second and third speakers were in a direction of ±30°, ±60° and ±90° with respect to the *SIG*. The loudspeakers simultaneously emitted voices of the same volume. The filter pass range $\delta(\theta)$ was ±20° when the loudspeaker was in the direction of 0° and 30°,

Metrics for evaluation		$R_1(dB)$				$R_2(dB)$				$R_3(dB)$			
Direction of a speaker		0°	30°	60°	90°	0°	30°	60°	90°	0°	30°	60°	90°
Condition of ADPF	**A**	2.0	1.3	2.2	0.5	-2.8	-3.1	-3.3	-7.7	10.4	4.7	2.6	-3.5
	B	2.2	1.4	1.6	0.8	-2.1	-3.4	-3.8	-7.3	9.1	4.6	3.4	-2.8
	C	2.2	1.1	2.1	0.6	-2.5	-4.0	-3.3	-7.7	10.3	6.8	2.6	-3.5

Table 1: Evaluation by R_1, R_2 and R_3 for separation of two simultaneous speeches

Interval of speakers		30°			60°			90°		
Direction of each speaker		−30°	0°	30°	−60°	0°	60°	−90°	0°	−90°
Condition of ADPF	**A**	4.9	8.1	5.1	3.2	9.6	3.1	-1.9	10.5	-1.7
	B	4.8	9.1	4.7	3.7	9.2	3.8	-1.7	9.1	-1.3
	C	5.7	7.4	5.9	3.5	9.6	3.1	-2.0	9.8	-2.0

Table 2: Evaluation by $R_3(dB)$ for separation of three simultaneous speeches

and was ±30° when the loudspeaker was in the direction of 60° and 90°. These values were defined according to the performance of the auditory fovea for a single sound source.

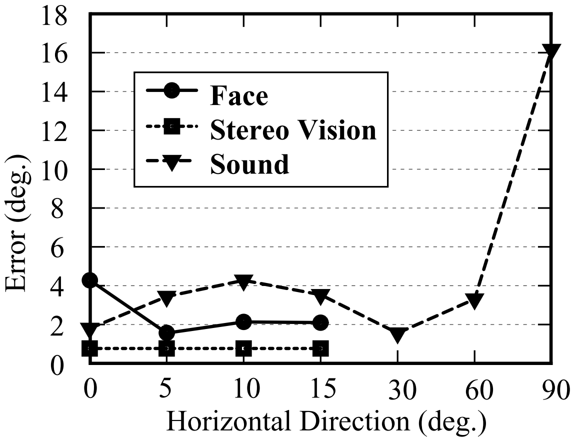

Figure 8: Error of localization by Face, Stereo Vision and Sound

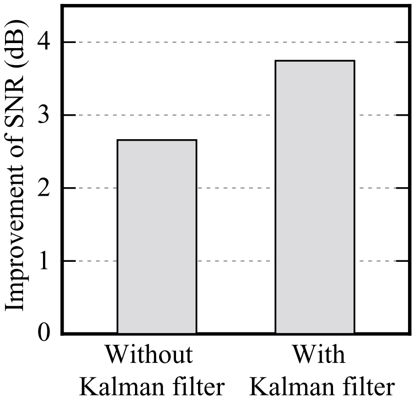

Figure 9: Moving speaker extraction

Sound source localization by Stereo Vision was the most accurate, as shown in Figure 8. The error was within 1°. Generally, localization by vision is more accurate than by audition. However, Sound has the advantage of an omni-directional sensor. That is, Sound can estimate the direction of sound from more than ±15° of azimuth. The sensitivity of localization by Sound depends on the sound source direction. It was the best in the front direction. The error was within ±5° from 0° to 30° and became worse at more than 30°. This proves that the correctness of using the auditory fovea and the efficiency of active motion, such as turning to face a sound source.

Figure 9 shows that the performance of the ADPF was increased by about 1 dB by the Kalman filter based stream formation. This indicates that the Kalman filter provides improved stream formation and accurate sound source direction.

Tables 1 and 2 show the results of sound source separation of two and three simultaneous voices, respectively. The similar tendencies of the performance were found in all filtering conditions. The difference between filtering condition **A**, which uses frequencies of below than 1500 Hz, and the other conditions is small, because sub-bands with frequencies higher than 1500 Hz collected by IID have lower power. This proves that auditory epipolar geometry is enough to separate sound sources by the ADPF even in real-world environments. R_1 and R_3 were the optimum in the front direction of *SIG*, and became worse towards the periphery. In the front direction, the efficiency of noise suppression was about 9 dB even with three simultaneous voices. But separation of two speakers closer together than 30° would be more difficult.

The loss of signal is 2–4 dB by R_2 in Table 1. According to two experts in audio signal processing, the filtering condition with the best clarity is **C**. The quality of the separated sounds is as good as the separation by 14 ch linear or 16 ch circle microphone arrays. On evaluation by listening, the ADPF is good at sound source separation. Our method of sound source separation currently does not have a function to cope with overlap in the frequency sub-bands of the differenct sources. This means that the system deals with sound mixture with a small amount of overlap such as voices properly. The system should have a function to solve overlap problems for more precise separation and separation of sounds with a lot of overlap such as music.

Conclusion

This paper reports sound source separation by an ADPF connected to a real-time multiple speaker tracking system. The ADPF with adaptive sensitivity control is shown to be effective in improving sound source separation. The sensitivity of the ADPF has not been reported so far in the literature and the idea of the ADPF lies in active motion to face a sound source to make the best use of the sensitivity. The ADPF would be efficient for front-end processing for Automatic Speech Recognition (ASR). The combination of the most up-to-date robust automatic speech recognition with the ADPF filter is one exciting research project now being planned.

Acknowledgements

We thank Prof. Tatsuya Kawahara of Kyoto University and Prof. Seiichi Nakagawa of Toyohashi University of Technology for their invaluable advice. We also thank Ken-ichi Hidai of Sony Corporation for his contribution which was made when he was in our laboratory and our colleagues of Symbiotic Intelligence Group, Kitano Symbiotic Systems Project for their discussions.

References

Aloimonos, Y.; Weiss, I.; and Bandyopadhyay., A. 1987. Active vision. *International Journal of Computer Vision*.

Asano, F.; Asoh, H.; and Matsui, T. 1999. Sound source localization and signal separation for office robot “jijo-2”. In *Proc. of IEEE Int. Conf. on Multisensor Fusion and Integration for Intelligent Systems (MFI-99)*, 243–248.

Asano, F.; Goto, M.; Itou, K.; and Asoh, H. Sep. 2001. Real-time sound source localization and separation system and its application to automatic speech recognition. In *Proceedings of International Conference on Speech Processing (Eurospeech 2001)*, 1013–1016. ESCA.

Blauert, J. 1999. *Spatial Hearing*. The MIT Press.

Faugeras, O. D. 1993. *Three Dimensional Computer Vision: A Geometric Viewpoint*. MA.: The MIT Press.

Hidai, K.; Mizoguchi, H.; Hiraoka, K.; Tanaka, M.; Shigehara, T.; and Mishima, T. 2000. Robust face detection against brightness fluctuation and size variation. In *Proc. of IEEE/RAS Int. Conf. on Intelligent Robots and Systems (IROS-2000)*, 1397–1384. IEEE.

Hiraoka, K.; Yoshizawa, S.; Hidai, K.; Hamahira, M.; Mizoguchi, H.; and Mishima, T. 2000. Convergence analysis of online linear discriminant analysis. In *Proc. of IEEE/INNS/ENNS Int. Joint Conference on Neural Networks*, III–387–391. IEEE.

Kagami, S.; Okada, K.; Inaba, M.; and Inoue, H. 1999. Real-time 3d optical flow generation system. In *Proc. of Int. Conf. on Multisensor Fusion and Integration for Intelligent Systems (MFI’99)*, 237–242.

Klarquist, W., and Bovik, A. 1998. Fovea: A foveated vergent active stereo vision system for dynamic 3-dimensional scene recovery. *RA* 14(5):755–770.

Nakadai, K.; Lourens, T.; Okuno, H. G.; and Kitano, H. 2000a. Active audition for humanoid. In *Proc. of 17th National Conference on Artificial Intelligence (AAAI-2000)*, 832–839. AAAI.

Nakadai, K.; Matsui, T.; Okuno, H. G.; and Kitano, H. 2000b. Active audition system and humanoid exterior design. In *Proc. of IEEE/RAS Int. Conf. on Intelligent Robots and Systems (IROS-2000)*, 1453–1461. IEEE.

Nakadai, K. Hidai, K.; Mizoguchi, H.; Okuno, H. G.; and Kitano, H. 2001. Real-time auditory and visual multiple-object tracking for robots. In *Proc. of the 17th Int. Joint Conf. on Atificial Intelligence (IJCAI-01)*, 1424–1432.

Nakagawa, Y.; Okuno, H. G.; and Kitano, H. 1999. Using vision to improve sound source separation. In *Proc. of 16th National Conference on Artificial Intelligence (AAAI-99)*, 768–775. AAAI.

Okuno, H.; Nakadai, K.; Lourens, T.; and Kitano, H. 2001. Separating three simultaneous speeches with two microphones by integrating auditory and visual processing. In *Proc. of European Conf. on Speech Processing(Eurospeech 2001)*. ESCA.

Okuno, H.; Ikeda, S.; and Nakatani, T. 1999. Combining independent component analysis and sound stream segregation. In *Proc. of IJCAI-99 Workshop on Computational Auditory Scene Analysis (CASA’99)*, 92–98. IJCAI.

Rougeaux, S., and Kuniyoshi, Y. 1997. Robust real-time tracking on an active vision head. In *Proc. of IEEE/RAS Int. Conf. on Intelligent Robots and Systems (IROS-97)*, 873–879. IEEE.

Schuller, G., and Pollak, G. Disproportionate frequency representation in the inferior colliculus of horsehoe bats: evidence for an “acoustic fovea”. In *J. Comp. Physiol. A*.